4-21-2013

Reliable Peak Selection for Multisample Analysis with Comprehensive Two-Dimensional Chromatography

Stephen E. Reichenbach
University of Nebraska-Lincoln, reich@cse.unl.edu

Xue Tian
University of Nebraska-Lincoln

Akwasi A. Boateng
USDA-ARS, Akwasi.Boateng@ars.usda.gov

Charles A. Mullen
USDA-ARS, charles.mullen@ars.usda.gov

Chiara Cordero
Universita` degli Studi di Torino

See next page for additional authors

analytical chemistry

Article

pubs.acs.org/ac

Reliable Peak Selection for Multisample Analysis with Comprehensive Two-Dimensional Chromatography

Stephen E. Reichenbach,*[,†] Xue Tian,[†] Akwasi A. Boateng,[‡] Charles A. Mullen,[‡] Chiara Cordero,*[,§] and Qingping Tao*[,∥]

[†]University of Nebraska − Lincoln, Lincoln, Nebraska 68588-0115, United States

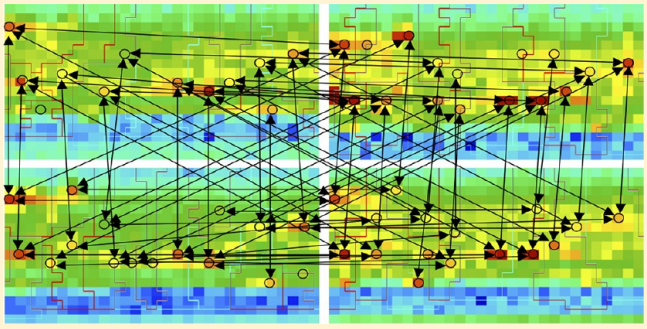
[‡]Sustainable Biofuels and Co-Products Research Unit, USDA-ARS, Eastern Regional Research Center, Wyndmoor, Pennsylvania 19038-8598, United States

[§]Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via P. Giuria 9, I-10125 Torino, Italy

[∥]GC Image, LLC, PO Box 57403, Lincoln, Nebraska 68505-7403, United States

Supporting Information

ABSTRACT: Comprehensive two-dimensional chromatography is a powerful technology for analyzing the patterns of constituent compounds in complex samples, but matching chromatographic features for comparative analysis across large sample sets is difficult. Various methods have been described for pairwise peak matching between two chromatograms, but the peaks indicated by these pairwise matches commonly are incomplete or inconsistent across many chromatograms. This paper describes a new, automated method for postprocessing the results of pairwise peak matching to address incomplete and inconsistent peak matches and thereby select chromatographic peaks that reliably correspond across many chromatograms. Reliably corresponding peaks can be used both for directly comparing relative compositions across large numbers of samples and for aligning chromatographic data for comprehensive comparative analyses. To select reliable features for a set of chromatograms, the Consistent Cliques Method (CCM) represents all peaks from all chromatograms and all pairwise peak matches in a graph, finds the maximal cliques, and then combines cliques with shared peaks to extract reliable features. The parameters of CCM are the minimum number of chromatograms with complete pairwise peak matches and the desired number of reliable peaks. A particular threshold for the minimum number of chromatograms with complete pairwise matches ensures that there are no conflicts among the pairwise matches for reliable peaks. Experimental results with samples of complex bio-oils analyzed by comprehensive two-dimensional gas chromatography (GCxGC) coupled with mass spectrometry (GCxGC−MS) indicate that CCM provides a good foundation for comparative analysis of complex chemical mixtures.



Multidimensional separation technologies, such as comprehensive two-dimensional gas chromatography (GCxGC) and comprehensive two-dimensional liquid chromatography (LCxLC), hold great promise for the analytical challenge of making comprehensive chemical comparisons across many samples. Comprehensive chemical comparisons provide the basis not only for sample comparisons but also for chemical fingerprinting, sample classification, chemical monitoring, sample clustering, and chemical marker discovery. However, matching chemical features for comparison across many complex samples is extremely difficult, and this difficulty remains a significant impediment to realizing the promise of multidimensional separations. This paper describes a new tool for one of the core problems in comparative analyses—a method for selecting chromatographic peaks that reliably correspond over large numbers of samples. This method, the Consistent Cliques Method (CCM), selects peaks that (a) match reliably across a large set of chromatograms and (b) have no inconsistencies in their pairwise matches.

Feature matching is the problem of establishing correspondences among attributes of different objects. In some pattern analysis problems, features or attributes are explicitly labeled so the correspondences are known and feature matching is not required. For example, if fish are to be classified on the basis of physical attributes such as weight and length, the values of those attributes are labeled such that one value is known to be the weight and another value is known to be the length. Then, weights are compared to weights and lengths are compared to lengths.

In other pattern analysis problems, features or attributes are not labeled and correspondences must be inferred. For

Received: January 17, 2013
Accepted: April 21, 2013
Published: April 21, 2013

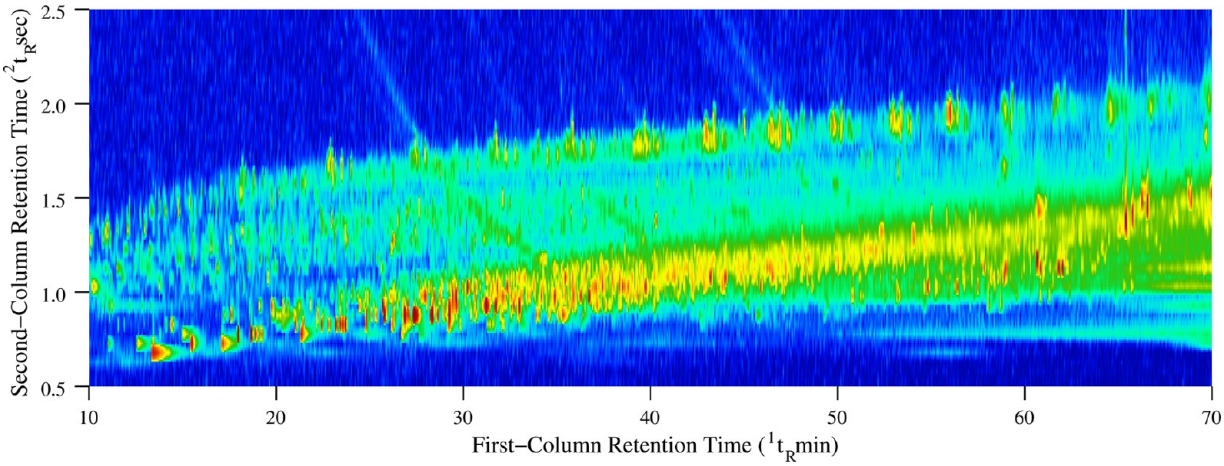


Figure 1. Pseudocolorized image of the total intensity count (TIC) for a GCxGC-MS chromatogram from a complex bio-oil. Each compound produces a two-dimensional peak in the data array output by the detector. Only a subregion is shown.

example, a classic problem requiring feature matching is image alignment.[1] Features such as edges or corners are detected in each image, but the correspondences, for example, which corner in one image matches to which corner in another image, are unknown. Feature matching establishes such correspondences and is important for a variety of tasks including analyzing, aligning, and comparing patterns.

The motivating application for this research is the analysis of large numbers of complex multidimensional patterns in data produced by GCxGC, LCxLC, or other multidimensional analytical approaches providing comprehensive information about analytes' properties. GCxGC separates complex mixtures using two columns interfaced by a modulator and connected to a detector.[2,3] If the chromatographic separation is fully effective, each compound produces a brief, resolved peak in the two-dimensional data. The GCxGC chromatogram of a complex mixture will exhibit hundreds or thousands of peaks, each of which is a characteristic feature of the data from that sample. Figure 1 illustrates the most relevant region of a GCxGC chromatogram of a complex bio-oil in which the x-axis is the first chromatographic column (^1D) retention time (^1t_R), the y-axis is the second chromatographic column (^2D) retention time (^2t_R), and the color indicates the relative value of the detector response. In this image, the value at each pixel is total intensity count (TIC) of the mass spectrum at the corresponding retention times and the pseudocolor is determined by a conventional cold-hot color map (in which the color progression blue, cyan, green, yellow, and red indicates increasing value), with automated value mapping to accentuate smaller peaks.[4]

A fundamental problem in GCxGC data analysis is to identify the compound that produces each peak—in other words, the labeling of each peak with its chemical identity. If the peak for a known compound is uniquely identified in each chromatogram, for example, by its retention times and/or spectrum, then that feature of the sample data can be labeled and compared directly across samples. However, even when the chemical identity for a peak cannot be established definitively, as is common for peaks in complex mixtures, comparative analysis requires that peaks be labeled (e.g., with an identification number) consistently across samples such that peaks resulting from the same compound in different samples have the same label (even if the compound identity is unknown). Therefore, comprehensive comparative analyses of complex samples by two-dimensional chromatography requires feature matching.

Figure 2 illustrates the problem of matching peak features for uniform labeling. In the top subimage of a chromatogram, there

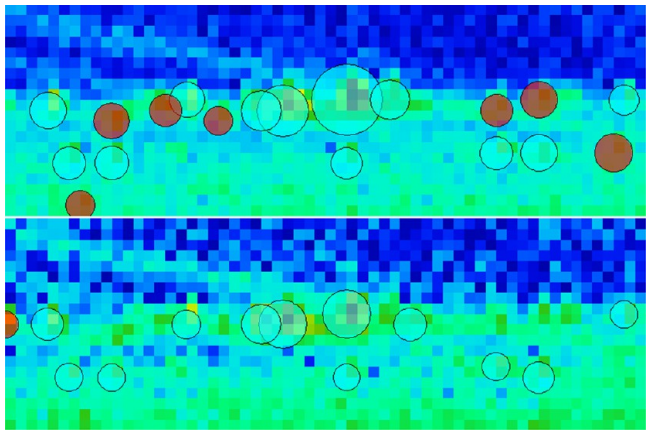


Figure 2. Small subregions of two chromatograms from different bio-oil samples. The data points in this figure are shown as rectangles to clearly illustrate the granularity of the modulator and detector sampling. The peaks marked by cyan-colored bubbles are definitively matched in each direction, but the peaks marked by red-colored bubbles are not definitively matched.

are nineteen overlaid semitransparent bubbles (some cyan and some red) indicating the locations and intensities (by bubble position and area) of detected peaks. In the bottom subimage of the same region in a different chromatogram, there are only thirteen detected peaks. Some of the peaks in the top subimage can be matched to peaks in the bottom subimage and vice versa. For example, the 12 prominent peaks with cyan-colored bubbles in each image can be matched to the 12 peaks with cyan-colored bubbles in the other image (even if the retention times and mass spectra of the peaks do not provide definitive compound identifications). However, matching of the other peaks with red-colored bubbles is not definitive, because of differences in the numbers of detected peaks, their retention times, and/or their mass spectra. Such differences may be due to compositional differences between samples, chromatographic variations, instrument noise, and/or data processing issues (e.g., peak detection errors).

dx.doi.org/10.1021/ac303773v | Anal. Chem. 2013, 85, 4974−4981

Various researchers have proposed alternative methods for pairwise peak matching.[5] This research does not address the problem of pairwise peak matching (nor of peak detection). Instead, this research addresses the problem of resolving incomplete and conflicting pairwise matches across many chromatograms. Accordingly, CCM can be used with any method for pairwise peak matching. Here, the template matching method[6−8] is used to pairwise match the pattern of peaks observed in one chromatogram (referred to as the template) to the peaks detected in another chromatogram (referred to as the target). Template matching uses both the retention-times pattern and spectral matching criteria. Template matching returns zero or one matching target peak for each template peak and each matched target peak is the best match for the matching template peak, subject to user-specified constraints and consistency with other peak matches. Alternative peak matching algorithms and/or different parameters might potentially improve pairwise matching performance, but unmatched and mismatched peaks are inevitable for large sets of complex chromatograms.

Consider the problem of finding reliable peaks that match not just for pairs of chromatograms, but across large sample sets of up to hundreds of chromatographic patterns. Figure 3

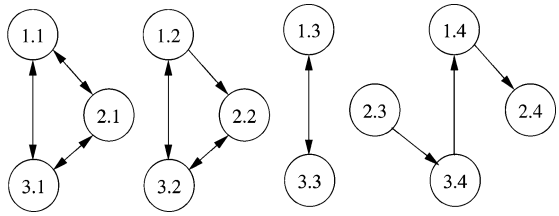


Figure 3. Example pairwise matchings, shown by arrows, between four peak features in each of three chromatographic patterns — Peaks 1.1−1.4 in Pattern 1, Peaks 2.1−2.4 in Pattern 2, and Peaks 3.1−3.4 in Pattern 3.

shows a graph that illustrates pairwise matchings between peaks in three chromatograms. In the graph, each peak is represented by a vertex, shown as a circle labeled with <chromatogramID>.<peakID>, and each pairwise template-to-target peak matching is indicated by a directed edge or arrow. The matchings for Peaks 1.1, 2.1, and 3.1 are reliable over every pair of chromatograms. For Peaks 1.2, 2.2, and 3.2, there are pairwise matchings between Chromatograms 1 and 3 and between Chromatograms 2 and 3, but there is only partial matching between Chromatograms 1 and 2 because Peak 2.2 failed to match Peak 1.2. The matchings for Peaks 1.3 and 3.3 are incomplete because no peak in Chromatogram 2 is matched. The matchings for 2.3, 1.4, 2.4, and 3.4 are conflicting. Over large numbers of complex samples, such partial, incomplete, and conflicting pairwise peak matches are common, so identifying peaks that reliably match across many chromatograms can be difficult.

Several possible approaches have been suggested to find peaks that match across multiple chromatograms. One approach is to determine reliable peaks by hand.[9,10] Such an approach may be able to achieve better success than automated methods, but is subjective and extremely tedious, potentially requiring days of manual labor.[10] Another approach is to designate a reference chromatogram and match all peaks in other chromatograms to it.[11−15] However, in many applications there is no true reference chromatogram and this approach could yield different results depending on the arbitrary selection of a reference chromatogram. Even if there is a natural choice for the reference chromatogram, that chromatogram may not exhibit peaks that could be reliably matched across many other chromatograms to provide important chemical information. Another approach is to proceed sequentially through the chromatograms, progressively modifying the set of peaks,[16,17] but such methods can yield different results depending on the ordering and some sets of chromatograms have no natural ordering.

The approach developed in this work, as described in the next section, is automated and does not bias the result by the selection of a reference chromatogram nor by the ordering of the chromatograms. This new approach automatically considers all pairwise matches in an objective manner to find peaks that can be matched reliably across the set of chromatograms. Once a set of reliably matched peaks is identified, they can be used for direct comparison or for other tasks such as alignment.

METHODS AND THEORY

Restrictive Algorithm and Concepts. A Restrictive Algorithm for finding reliable peaks, shown in Figure 4, selects peaks that have complete pairwise matches across every chromatogram—that is, peaks for which there are no partial matches, incomplete matches, or conflicts in any of the pairwise matches across all chromatograms. The Restrictive Algorithm does this by finding cliques that are as large as the number of chromatograms. A *clique* is a subset of the vertices of a graph such that every vertex in the subset is connected to every other vertex in the subset; that is, every pair of chromatographic peaks in a clique are pairwise matched with one another. For example, in the graph of Figure 3, the set of peaks {1.1, 2.1, 3.1} form a clique across all three chromatograms because there is a match between every pair of template and target peaks in the set: (1.1 → 2.1), (1.1 → 3.1), (2.1 → 1.1), (2.1 → 3.1), (3.1 → 1.1), and (3.1 → 2.1). However, the set of peaks {1.2, 2.2, 3.2} do not form a clique across all three chromatograms because it is not true that there is a match between every pair of template and target peaks: specifically, Peak 2.2 does not match Peak 1.2.

Restrictive Algorithm:

1. Detect the peaks in each chromatogram.
2. Perform all pairwise peak matchings between chromatograms.
3. Build a graph, as in Figure 3, in which each peak from Step 1 is a vertex and each matching of a template peak in one chromatogram to a target peak in another chromatogram from Step 2 is a directed edge.
4. Find and report cliques in the graph from Step 3 that contain a peak from each chromatogram.

Figure 4. Restrictive Algorithm.

dx.doi.org/10.1021/ac303773v | *Anal. Chem.* 2013, 85, 4974−4981

CCM Algorithm:

1. Build a graph from pairwise peak mathings by performing Steps 1–3 of the Restrictive Algorithm.
2. Find maximal cliques in the graph from Step 1 that contain peaks in at least s chromatograms, where s is a parameter of the algorithm.
3. Combine the maximal cliques from Step 2 that share common peaks.
4. Report the combined cliques from Step 3 in order from largest to smallest until the number of reliable peak features requested by the user are reported or all combined cliques are reported.

Figure 5. CCM Algorithm.

Each clique reported in Step 4 contains a set of peaks that are pairwise matched across all chromatograms. Peaks that do not match pairwise across all chromatograms are not reported as reliable features. The Supporting Information describes some of the properties of these matching graphs in more detail.

Unfortunately, this Restrictive Algorithm has two problems. First, this approach does not find peak features that are in most chromatograms but which are undetected in at least one chromatogram. If the goal is alignment, then this issue may not be a problem as long as enough reliable peaks are identified; but, if the goal is comparison, then some pertinent peak features may not be selected for comparison. The second issue is more serious: this approach does not scale well for large sample sets. If the peak matching is relatively reliable but imperfect (as most real-world phenomena are), then as the number of chromatograms increases, the likelihood of a failed match or inconsistency—even for highly reliable features—grows exponentially.

To illustrate this second problem, note that the number of possible features with a matched peak in each chromatogram is limited by the number of peaks in the chromatogram with the fewest detected peaks. On the other hand, the number of pairwise matches required for a feature with a peak in each chromatogram is equal to $n(n-1)$, where n is the number of chromatograms, because a clique must have n matching peaks and each peak must match to a peak in each of the other $(n-1)$ chromatograms. So, while the number of possible features is fixed or diminishes as new chromatograms are acquired, the requirements for complete consistency for any feature increases exponentially with larger numbers of chromatograms.

For example, if pairwise peak matches are 99.9% reliable, then for a set of ten chromatograms, more than 91% of such peaks are expected to be matched across all chromatograms $(0.999^{10(10-1)} = 0.913890)$; but for a set of 100 chromatograms, fewer than one in 20000 of such peaks are expected to be matched across all chromatograms $(0.999^{100(100-1)} = 0.000499)$. In this example of 100 chromatograms, if the chromatogram with the fewest peaks contains only a few thousand peaks, then it is likely that no reliable peaks would be found. For peaks with lower pairwise-matching reliability, the problem is apparent for even smaller sets of chromatograms. As described in the next subsection, the solution for these issues is to allow the user to relax the requirement for complete pairwise matching across all chromatograms.

Consistent Cliques Method. The CCM, developed in this paper, is shown in Figure 5. Step 1 of CCM is the same as the Restrictive Algorithm through Step 3, but CCM then selects peaks that are consistently matched over some, but perhaps not all, chromatograms. Step 2 finds cliques that contain peaks from at least s chromatograms, but which do not necessarily contain peaks from all chromatograms. The user can reduce the minimum size of the maximal cliques, s in Step 2, thereby yielding additional but less reliable peaks with pairwise matches in fewer chromatograms. The percentage of chromatograms that have peaks in the clique can be regarded as a measure of the reliability of a peak feature with respect to a set of chromatograms. For example, if a peak is matched consistently across 12 of 15 chromatograms, it can be said to be 80% reliable.

Cliques that are smaller than the number of chromatograms may share common peaks, that is, peaks that should be regarded as one common feature could form more than one maximal clique. So, Step 3 combines those cliques sharing common peaks. The resulting combined cliques may be of different sizes, so, in Step 4, if the user asks for a number of reliable features that is less than the number of combined cliques, then only the most reliable features with peaks in the largest number of chromatograms are reported, up to the number of requested features.

For example, given the graph in Figure 3, CCM Step 2 detects the four maximal cliques of size $s = 2$ or larger: {1.1, 2.1, 3.1}, {1.2, 3.2}, {2.2, 3.2}, and {1.3, 3.3}. In Step 3, cliques {1.2, 3.2} and {2.2, 3.2} are combined to form {1.2, 2.2, 3.2}, because they have Peak 3.2 in common. In this way, CCM finds the peak feature that was missed by the Restrictive Algorithm. If the user asks for two features, then only {1.1, 2.1, 3.1} and {1.2, 2.2, 3.2} are reported; but if the user asks for more than two features, then {1.3, 3.3} also is reported.

Setting $s > \lceil 2n/3 \rceil$, where n is the number of chromatograms, ensures that sets of feature cliques that share common peaks are conflict-free; that is, the union of such cliques has no more than one peak in each chromatogram. The proof of this is provided in the Supporting Information. If there are no conflicts, then feature cliques that share common peaks in Step 3 can be combined by a simple union. The minimum size of the maximal feature cliques can be fixed to the smallest value that ensures conflict-free results:

$$s_n = \lceil (2n+1)/3 \rceil \quad (1)$$

The user still is provided parametric control of the number of desired features, in Step 4, to constrain the relative reliability of the reported features.

With $s = s_n$ in Step 2 and the union of cliques in Step 3, CCM selects peaks that are consistently matched across more than two-thirds of the chromatograms. The threshold s can be set to a smaller number, but then it might be necessary to deal with conflicts between cliques that share common peaks in Step 3. Three possible alternative methods for dealing with such conflicts in the combining of cliques that share common peaks are: (a) eliminate from the combination those peaks in the chromatograms for which there is a conflict, (b) eliminate from the combination all cliques for which there is a conflict, and (c)

do not report a combination of cliques for which there is a conflict.

EXPERIMENTAL PROCEDURES

Example Analysis. The CCM for selecting reliable peaks is demonstrated here with GCxGC-MS analyses of complex upgraded pyrolysis oils from the presscake of pennycress seeds (*Thlaspi arvense* L.). Pennycress makes a good bioenergy crop because it is a winter crop and therefore can be an additional crop that does not replace a food crop. The presscake is the material remaining after mechanical pressing to remove most of the vegetable oil (which is used for biodiesel or green diesel production). The presscake is pyrolyzed to produce bio-oil.

Fast pyrolysis oils from biomass materials with high protein content generally are more stable and partially deoxygenated compared with those from mostly lignocellulosic biomass (e.g., wood, grasses) due to nucleophilic substitution of nitrogen for oxygen. However, in order for these products to be used as transportation fuels or petroleum refinery feedstocks, the pyrolysis oils still must be upgraded to reduce their heteroatom (N, O, S) content. Because the compositions of these proteinaceous pyrolysis oils differ greatly from those from lignocellulosic feedstocks, their behavior in various upgrading steps will be different. Therefore, research on Sustainable Biofuels and Coproducts at the Eastern Regional Research Center, Agricultural Research Service (ARS), U.S. Department of Agriculture (USDA), is studying hydrotreating fast pyrolysis oils from the presscake of pennycress seeds as a model for upgrading proteinaceous fast pyrolysis oils.[18]

The goals for chemical analysis include characterizing the chemical transformations that occur, comparing these products with those from hydrogenation of lignocellulosic pyrolysis oils, and comparing the selectivity of the catalysts for some of the individual reactions in the complex system. Here, however, the experiments are used only to demonstrate the CCM method for selecting reliable peaks.

Sample Production, Analysis, and Data Processing. Three experimental replicates (i.e., from different samples under the same conditions) for each of five catalysts were produced at the Sustainable Biofuels and Co-Products Research Unit, USDA-ARS, as described in the Supporting Information.

Each sample was analyzed by GCxGC-MS at the USDA-ARS using a Shimadzu (Kyoto, JP) GCMS-QP2010S GC-MS system and a Zoex (Houston TX, USA) ZX-2 LN2 cooled-loop GCxGC thermal modulation system, as described in the Supporting Information. The resulting 15 chromatograms are pictured in the Supporting Information.

Automated data processing was performed at the University of Nebraska − Lincoln, using GC Image (Lincoln NE, USA) GCxGC Edition Software, R2.3a0.

1. Preprocessing and peak detection. In each chromatogram, the data was shifted as necessary to align the first data-point relative to the modulation start-time.[8] Then, the baseline was corrected so that the peaks rise above a near-zero-mean baseline.[19] Then, the two-dimensional blob-peaks were detected using the Drain Algorithm,[7] which automatically performs true two-dimensional peak detection.[20]
2. Template construction and matching. From each chromatogram, a template[6,7] was constructed to record the retention times and normalized mass spectrum of each detected peak. Each template peak also was given a mass-spectral matching constraint written in CLIC,[21] which generally required that the NIST match factor[22] for matched peaks be at least 700 (although lower match factors were allowed for peaks if no nearby peak exhibited a similar spectrum). Then, the template from each chromatogram was matched[8] with the detected peaks for each of the other fourteen processed chromatograms, for a total of 210 pairwise template-to-target matches between chromatograms.

The next section examines the results for CCM operating on the pairwise peak-matching results generated by template matching.

RESULTS AND DISCUSSION

Peak Detection and Matching. This subsection describes the results of peak detection and matching outlined in the Experimental Procedures section. Again, note that CCM could use the output of any methods for peak detection and pairwise peak matching.

The average number of peaks detected in each bio-oil chromatogram was 567, with a range of 436 to 699 over the fifteen chromatograms. The average number of peaks in the chromatograms for each of the catalysts did not vary greatly, with 571, 515, 608, 587, and 552 peaks respectively for Catalysts 1 to 5.

Table 1 shows the average peak-matching rates between chromatograms for the five catalysts. The overall average peak-matching rate between pairs of chromatograms was 41.9%, which is a fairly low rate, reflecting compositional differences; the large dynamic range of peak intensities, including many faint peaks; and peak crowding, which complicates with peak detection. The average peak-matching rate between chromatograms for the same catalyst (six pairwise matches each) was higher than the overall average, at 43.6%. By comparison, the average peak-matching rate between chromatograms for different catalysts (eighteen pairwise matches each) was 41.5%. Even if the pairwise matching could be improved by better tuning of the template-matching parameters or by using another peak-matching method, these numbers indicate that, for this data, pairwise matching is challenging and that across the such large, complex sets of chromatograms there will be many incomplete and inconsistent matches.

Table 1. Percentage Peak-Matching Rates by Catalyst

	template catalyst					
target catalyst	1	2	3	4	5	average
1	43.9	43.0	39.2	43.1	42.2	42.3
2	38.9	41.7	34.0	37.2	43.0	38.9
3	41.5	40.0	43.8	45.6	41.0	42.4
4	44.3	42.2	44.2	42.8	44.2	43.5
5	41.1	46.2	37.6	41.9	45.9	42.6
average	41.9	42.6	39.8	42.1	43.2	41.9

Figure 6 shows a small subset of the pairwise matches across a subset of the chromatograms. The regions shown contain less than 3% of the detected peaks in the full chromatograms and the number of pairwise matches among these 4 chromatograms is less than 6% of the number of pairwise matches across the full set of 15 chromatograms. The graph for this small subset of peaks in a small subset of chromatograms is quite complex, but the graph for the set of 15 complete chromatograms has nearly

dx.doi.org/10.1021/ac303773v | Anal. Chem. 2013, 85, 4974−4981

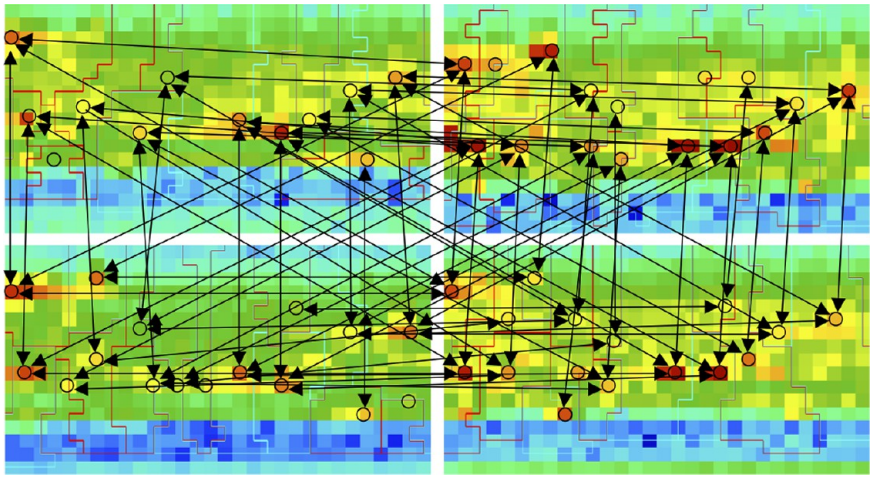


Figure 6. Detected peaks (shown with open circles) and pairwise matches (shown with directed arrows) in a small region in a subset of four of the bio-oil chromatograms.

40 times as many peaks and nearly 200 times as many matching edges. From such complex matching graphs, CCM automatically finds reliable peaks without biasing the search with a reference chromatogram or a specific chromatographic ordering.

Reliable Peaks. The Restrictive Algorithm for selecting reliable peaks found only 2 maximal cliques of size 15, indicating only 2 reliable peaks with complete pairwise matchings across all chromatograms. Two peaks are not sufficient to effectively characterize or compare the chromatograms nor even to determine a fully parametrized affine transformation for alignment. This example illustrates the need for a more flexible method for selecting reliable peaks and the motivation for CCM.

Figure 7 illustrates the number of maximal cliques as a function of the minimum clique size s. As just noted, there are

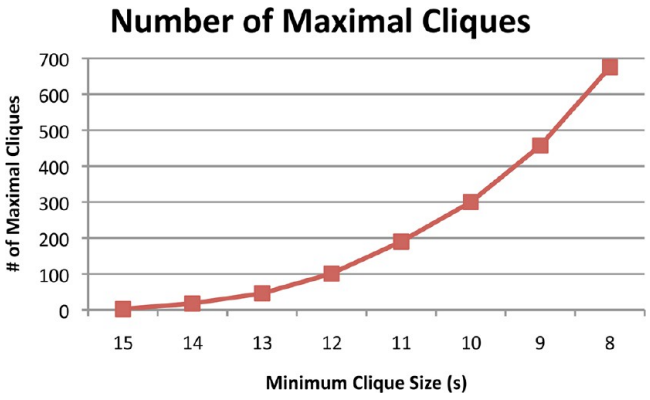


Figure 7. Graph shows the number of maximal cliques of peaks in the bio-oil chromatograms as a function of the mimimum clique size.

only two cliques with a peak in every one of the 15 chromatograms. However, as the minimum size is decreased, the number of maximal cliques that are sufficiently large increases. With the threshold for the clique size set as $s_{15} = 11$, which is the smallest threshold that guarantees that cliques that share a common peak are conflict-free, there are 190 maximal cliques. With the threshold for the clique size set as $s = 8$, which yields cliques that have complete pairwise matchings for more than half the chromatograms, there are 675 maximal cliques. It is possible to have more cliques than peaks because cliques may share common peaks.

After combining the 190 maximal cliques of size $s_{15} = 11$ or larger that share peaks, there are 80 combined cliques. These combined cliques indicate 80 reliable peaks matched in 11 or more chromatograms. The user can take a subset of these combined cliques to get only the more reliable peaks. For example, as shown by the red line with squares in Figure 8, 29 of the combined cliques have a peak in each of the 15 chromatograms, 45 of the combined cliques are size 14 or larger, etc.

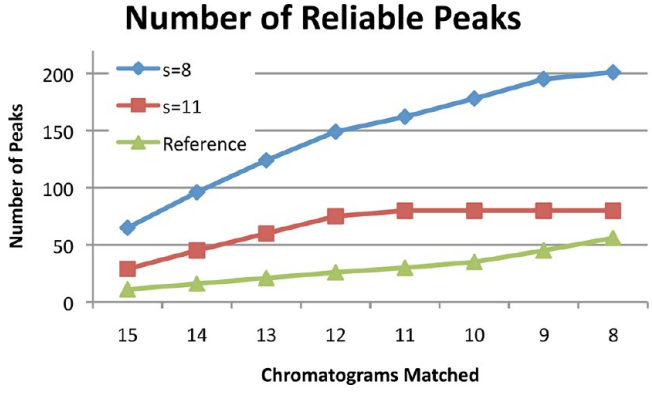


Figure 8. Graph shows the number of conflict-free, reliable peaks as a function of the number of chromatograms matched for (a) CCM with $s = 8$, (b) CCM with $s_{15} = 11$, and (c) using a reference chromatogram. At all levels of reliability, CCM yields a greater number of reliable peaks than does matching with a reference chromatogram.

After combining the 675 maximal cliques of size $s = 8$ or larger that share common peaks, there are 201 combined cliques, indicating 201 reliable peaks matched in eight or more chromatograms. With $s = 8$, conflicts for combined cliques are possible, but for this data, none of the combined cliques exhibit any conflict. As shown by the blue line with diamonds in Figure 8, combining maximal cliques of size $s = 8$ or larger yields 65 combined cliques with a peak in each of the 15 patterns, 96 are size 14 or larger, etc.

For a comparison with a previous approach to selecting reliable peaks, if the first chromatogram is used as a reference,

dx.doi.org/10.1021/ac303773v | Anal. Chem. 2013, 85, 4974−4981

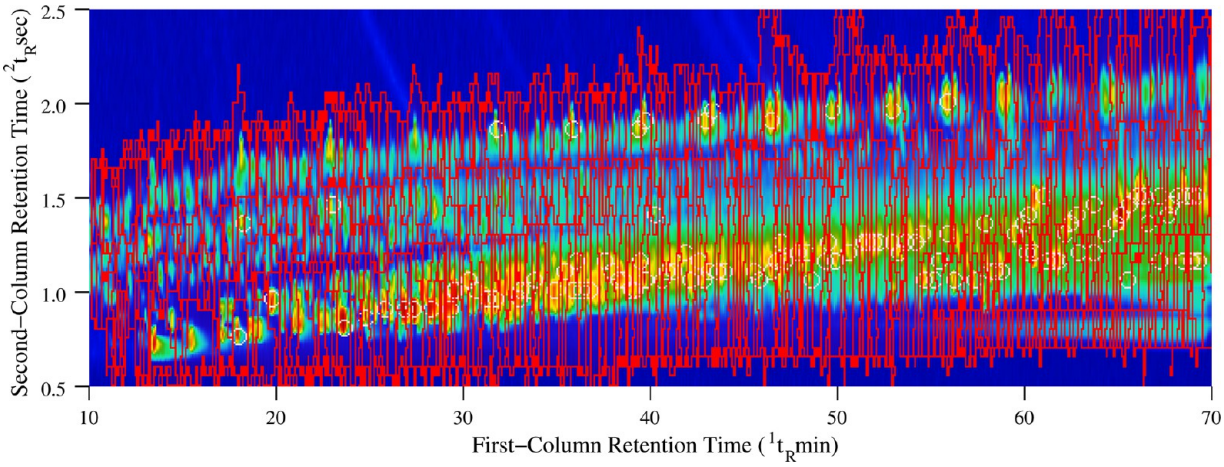


Figure 9. Pseudocolorized image of the total intensity count (TIC) for the composite chromatogram from fifteen aligned chromatograms. The overlay shows the reliable peaks (with white circles) and the regions of the detected peaks (with red dotted outlines).

11 of its peaks can be matched unidirectionally to a peak in every other chromatogram without matching conflicts among the 15 chromatograms. By comparison, CCM identifies many more conflict-free peaks that can be found in all 15 chromatograms: 29 for CCM with s_{15} = 11 and 65 for s = 8. As illustrated by the green line with triangles in Figure 8, if the number of chromatograms that must be matched without pairwise conflicts is reduced, the number of peaks that are reliable at that level increases. Reference matching involving 11 chromatograms (i.e., the reference chromatogram and ten others) yields 30 reliable peaks, compared with 80 reliable peaks from CCM with s_{15} = 11 and 162 reliable peaks from CCM with s = 8. Reference matching involving eight chromatograms yields 56 reliable peaks, compared with 201 reliable peaks from CCM with s = 8. CCM will identify all reliable peaks identified by the reference chromatogram method except those for which none of the matched peaks matches back to the reference peak, which is unlikely in practice and does not occur in this example. Additionally, CCM will identify peaks that are found in many chromatograms but not the reference chromatogram—peaks which the reference chromatogram method will not identify. In this example, at all levels of reliability, CCM yields a significantly larger number of reliable peaks than does matching with a reference chromatogram.

The ability of CCM to select reliable peaks is determined by the degree of reliable peak-matching. Consistent chromatographic performance and conditions that produce well resolved peaks will allow reliable pairwise peak-matching and so CCM could select reliable peaks at a high rate. Conversely, inconsistent conditions and chromatograms with poorly resolved peaks make pairwise peak-matching less reliable and so CCM would select reliable peaks at a lower rate. As noted in Peak Detection and Matching, the pairwise peak-matching rate for this data was fairly low—less than 44%—due to compositional differences, the large dynamic range of peak intensities, and peak crowding. Given the relatively low pairwise peak-matching rate for this data, CCM selected reliable peaks at a comparatively high rate—more than 35% of the average number of peaks in each chromatogram.

The computation time for the CCM algorithm to select the reliable peaks based on the input pairwise template matches is relatively small. For the 15 samples presented here, executing the CCM program required less than 30 s on a desktop personal computer.

Multisample Comparisons. The goal of this paper is to develop a new method for selecting peaks that are reliably matched across many chromatograms. This is one step in the larger process of comparative analysis. For the bio-oils example, comparative analysis would entail extracting and applying comprehensive chemical information in order to develop knowledge about the performance of the different catalysts in the pyrolysis of the bio-oils. This larger problem is beyond the scope of this paper and is addressed in other publications,[23] but a few comments about using reliable peaks for multisample comparisons are appropriate.

Reliable peaks can be used to directly compare chromatograms with respect to those indicated features and/or to align chromatograms in the retention-time plane for comprehensive comparisons. Figure 9 shows a composite chromatogram generated by aligning the fifteen bio-oil chromatograms, using an affine transformation that minimizes the mean-square misalignment for the matched reliable peaks. The overlay in Figure 9 shows the 201 reliable peaks from CCM with s = 8 using white circles and outlines the peaks detected in the composite chromatogram using red dotted lines. The reliable peaks selected by CCM can be listed in a table, each with concentration, compound identification, and/or other measures, either on a per-run basis (for chromatograms in which matches exist) or on an aggregate basis (e.g., by sample, class, or overall).

There are 660 peaks detected in the composite chromatogram, which is within the range of the number of peaks detected in the individual chromatograms. Note that this is more than 3 times the number of reliable peaks found by CCM with s = 8 and more than 11 times the number of reliable peaks from reference matching involving 8 chromatograms. For large numbers of such complex chromatograms, even tedious and time-consuming interactive selections by expert analysts cannot objectively match all peaks. For this reason, even with a method for relatively effective peak matching, peak-based approaches may not provide a basis for automating truly comprehensive nontargeted comparisons of complex samples. Region-based approaches, for example, using the retention-time windows defined by the detected peaks shown by red outlines in Figure 9, provide a better basis for automated, comprehensive, nontargeted comparisons of complex samples.[5] However,

dx.doi.org/10.1021/ac303773v | Anal. Chem. 2013, 85, 4974−4981

region-based approaches require chromatographic alignment, typically by aligning matched peaks, so methods for selecting reliably matching peaks still have an important role.

CONCLUSION

The CCM is an algorithm for selecting features that are reliably matched across many patterns. CCM can be used with applications that involve complex data with unlabeled features, such as comprehensive two-dimensional chromatography with unlabeled peaks. Comparative multisample analyses with comprehensive two-dimensional chromatography require peaks that are reliably matched (i.e., deemed to result from the same compound) across many samples. Such analyses may involve classification of samples, chemical fingerprinting, chemical monitoring, sample clustering, and chemical marker discovery. CCM overcomes problems with previous approaches: CCM is fully automated, it is not dependent on the selection of a reference pattern, and the result does not depend on the ordering of the patterns.

Here, CCM was demonstrated with 15 chromatograms of complex bio-oil samples with nearly 600 peaks detected, on average, for each sample. Only 2 peaks matched consistently across every one of the 210 possible pairwise matchings, but CCM identified more than 200 peaks that were matched across more than half of the chromatograms. The reliable peaks can be used to directly compare samples and/or to align the chromatograms for truly comprehensive comparisons, for example, with peak-region features.

Future work on CCM might relax the constraint that, in each pairwise matching, each feature is matched only with its best match. However, allowing more than one potential match in pairwise matchings could significantly increase computational complexity. Taken in its general form, the problem of whether a graph contains a clique larger than a given size is NP-complete, meaning that even moderately sized problems can be intractable. Moreover, such an approach could produce conflicting feature sets. In its current form, CCM can be performed rapidly, and if the minimum size of the clique is set to more than 2/3 of the number of samples, then the feature sets will be conflict-free.

ASSOCIATED CONTENT

Supporting Information

Supplementary figures. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: reich@cse.unl.edu, chiara.cordero@unito.it, qtao@gcimage.com.

Notes

The authors declare the following competing financial interest(s): Qingping Tao and Stephen E. Reichenbach have employment and financial interests in GC Image, LLC, which makes and sells software used for some of the data visualization, processing, and analysis.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. National Science Foundation under Award Number IIP-1013180 and by the Nebraska Center for Energy Sciences Research.

REFERENCES

(1) Horaud, R.; Skordas, T. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 1168−1180.
(2) Bushey, M. M.; Jorgenson, J. W. *Anal. Chem.* **1990**, *62*, 161−167.
(3) Liu, Z. Y.; Phillips, J. B. *J. Chromatogr. Sci.* **1991**, *29*, 227−331.
(4) Visvanathan, A.; Reichenbach, S. E.; Tao, Q. *J. Electron. Imaging* **2007**, *16*, 033004.
(5) Reichenbach, S. E.; Tian, X.; Cordero, C.; Tao, Q. *J. Chromatogr. A* **2012**, *1226*, 140−148.
(6) Reichenbach, S. E.; Carr, P. W.; Stoll, D. R.; Tao, Q. *J. Chromatogr. A* **2009**, *1216*, 3458−3466.
(7) Reichenbach, S. E.; Ni, M.; Kottapalli, V.; Visvanathan, A. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 107−120.
(8) Reichenbach, S. E. In *Comprehensive Two Dimensional Gas Chromatography*; Ramos, L., Ed.; Elsevier: Oxford UK, 2009; Chapter 4, pp 77−106.
(9) van Mispelaar, V. G. Chromametrics. Ph.D. thesis, University of Amsterdam, 2005.
(10) Koek, M. M.; van der Kloet, F. M.; Kleemann, R.; Kooistra, T.; Verheij, E. R.; Hankemeier, T. *Metabolomics* **2011**, *7*, 1−14.
(11) Shellie, R. A.; Welthagen, W.; Zrostliková, J.; Spranger, J.; Ristow, M.; Fiehn, O.; Zimmermann, R. *J. Chromatogr. A* **2005**, *1086*, 83−90.
(12) Wardlaw, G. D.; Arey, J. S.; Reddy, C. M.; Nelson, R. K.; Ventura, G. T.; Valentine, D. L. *Environ. Sci. Technol.* **2008**, *42*, 7166−7173.
(13) Oh, C.; Huang, X.; Regnier, F. E.; Buck, C.; Zhang, X. *J. Chromatogr. A* **2008**, *1179*, 205−215.
(14) Gaquerel, E.; Weinhold, A.; Baldwin, I. T. *Plant Physiol.* **2009**, *149*, 1408−1423.
(15) Li, X.; Xu, Z.; Lu, X.; Yang, X.; Yin, P.; Kong, H.; Yu, Y.; Xu, G. *Anal. Chim. Acta* **2009**, *633*, 257−262.
(16) Cordero, C.; Liberto, E.; Bicchi, C.; Rubiolo, P.; Schieberle, P.; Reichenbach, S. E.; Tao, Q. *J. Chromatogr. A* **2010**, *1217*, 5848−5858.
(17) Castillo, S.; Mattila, I.; Miettinen, J.; Orešič, M.; Hyötyläinen, T. *Anal. Chem.* **2011**, *83*, 3058−3067.
(18) Mullen, C. A.; Boateng, A. A. *BioEnergy Res.* **2011**, *4*, 303−311.
(19) Reichenbach, S. E.; Ni, M.; Zhang, D.; Ledford, E. B., Jr. *J. Chromatogr. A* **2003**, *985*, 47−56.
(20) Latha, I.; Reichenbach, S. E.; Tao, Q. *J. Chromatogr. A* **2011**, *1218*, 6792−6798.
(21) Reichenbach, S. E.; Kottapalli, V.; Ni, M.; Visvanathan, A. *J. Chromatogr. A* **2004**, *1071*, 263−269.
(22) Stein, S. E. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 770−781.
(23) Mullen, C. A.; Boateng, A. A.; Reichenbach, S. E. *Fuel* **2013**, DOI: 10.1016/j.fuel.2013.04.075.